Linear Regression Modelling of Abalone Dataset

Assessment 2 Report of ZZSC6836 - Data Mining and Machine Learning

Rushmila Islam

2023-11-05

Contents

1 Introduction

The Abalone dataset is a machine learning dataset available at the UCI Machine Learning Repository. The dataset contains measurements of abalones, which are large sea snails. The measurements include the length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight of each abalone. It is a popular choice for machine learning tasks such as regression and classification. The target variable is the age of the abalone, which is determined by counting the number of rings in its shell. The dataset is challenging because the age of an abalone is difficult to predict from its physical measurements. However, the dataset is well-suited for machine learning algorithms that can learn complex relationships between features. In this project, we perform exploratory data analysis of the Abalone dataset and then perform simple linear regression to predict the age of the abalone by varying the input features and applying different techniques. Our results suggest that determining an appropriate set of input features for such regression model is challenging and indicates more elaborate dataset containing additional features might have been helpful in determining the abalone age.

1.1 Background

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem. The dataset used in this analysis is obtained from the UCI Machine Learning Repository (UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science), at the following link: https://archive.ics.uci.edu/ml/datasets/abalone

1.1.1 Data Pre-processing

In this section, we perform basic data pre-processing tasks to check for missing values, renaming columns, and converting the categorical column values to numerical ones. We also observe the standard data descriptions like range, min-and-max values, percentile, mean, standard deviations, and the overall spread and variety of the individual column values.

Here is a column-level view of the cleaned and pre-processed Abalone dataset.

```
abalone.head()
```

```
##    Sex  Length  Diameter  Height  ...  Viscera_weight  Shell_weight  Rings  Ring-age
## 0    0   0.455     0.365   0.095  ...          0.1010         0.150     15      16.5
## 1    0   0.350     0.265   0.090  ...          0.0485         0.070      7       8.5
## 2    1   0.530     0.420   0.135  ...          0.1415         0.210      9      10.5
## 3    0   0.440     0.365   0.125  ...          0.1140         0.155     10      11.5
## 4    2   0.330     0.255   0.080  ...          0.0395         0.055      7       8.5
##
## [5 rows x 10 columns]
```

```
abalone.info()
```

```
## <class 'pandas.core.frame.DataFrame'>
## RangeIndex: 4177 entries, 0 to 4176
## Data columns (total 10 columns):
##  #   Column          Non-Null Count  Dtype
## ---  ------          --------------  -----
##  0   Sex             4177 non-null   int64
##  1   Length          4177 non-null   float64
##  2   Diameter        4177 non-null   float64
##  3   Height          4177 non-null   float64
##  4   Whole_weight    4177 non-null   float64
##  5   Shucked_weight  4177 non-null   float64
##  6   Viscera_weight  4177 non-null   float64
##  7   Shell_weight    4177 non-null   float64
##  8   Rings           4177 non-null   int64
##  9   Ring-age        4177 non-null   float64
## dtypes: float64(8), int64(2)
## memory usage: 326.5 KB
```

```
abalone.describe()
```

```
##                Sex       Length  ...        Rings     Ring-age
## count  4177.000000  4177.000000  ...  4177.000000  4177.000000
## mean      0.955470     0.523992  ...     9.933684    11.433684
## std       0.827815     0.120093  ...     3.224169     3.224169
## min       0.000000     0.075000  ...     1.000000     2.500000
## 25%       0.000000     0.450000  ...     8.000000     9.500000
## 50%       1.000000     0.545000  ...     9.000000    10.500000
## 75%       2.000000     0.615000  ...    11.000000    12.500000
## max       2.000000     0.815000  ...    29.000000    30.500000
##
## [8 rows x 10 columns]
```

1.1.2 Preparing dataset for modelling

In preparation for linear regression modelling, we separate out the target variable i.e., Ring-age and consider the rest of the columns as the input variables i.e., features for machine learning.

```
## Sex
## 0    1528
## 2    1342
## 1    1307
## Name: count, dtype: int64
```

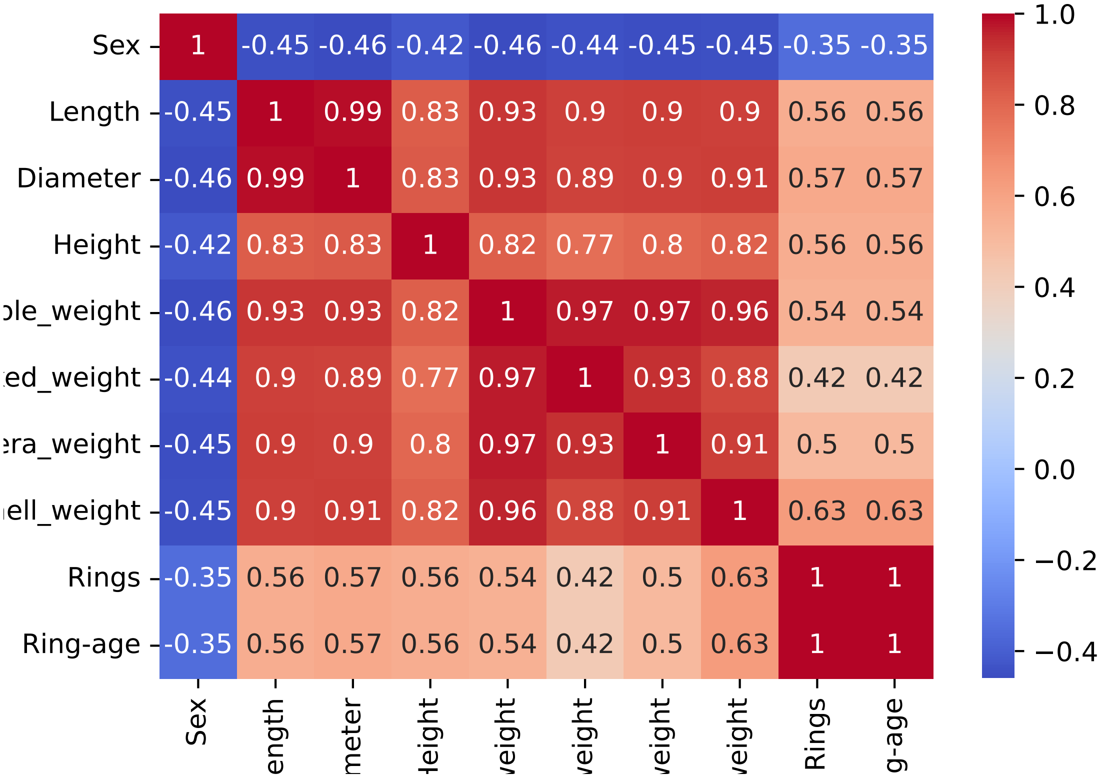
```
## <class ’pandas.core.frame.DataFrame’>
## RangeIndex: 4177 entries, 0 to 4176
## Data columns (total 9 columns):
##  #   Column          Non-Null Count  Dtype
## ---  ------          --------------  -----
##  0   Sex             4177 non-null   int64
##  1   Length          4177 non-null   float64
##  2   Diameter        4177 non-null   float64
##  3   Height          4177 non-null   float64
##  4   Whole_weight    4177 non-null   float64
##  5   Shucked_weight  4177 non-null   float64
##  6   Viscera_weight  4177 non-null   float64
##  7   Shell_weight    4177 non-null   float64
##  8   Rings           4177 non-null   int64
## dtypes: float64(7), int64(2)
## memory usage: 293.8 KB
```

```
## <class ’pandas.core.frame.DataFrame’>
## RangeIndex: 4177 entries, 0 to 4176
## Data columns (total 1 columns):
##  #   Column    Non-Null Count  Dtype
## ---  ------    --------------  -----
##  0   Ring-age  4177 non-null   float64
## dtypes: float64(1)
## memory usage: 32.8 KB
```

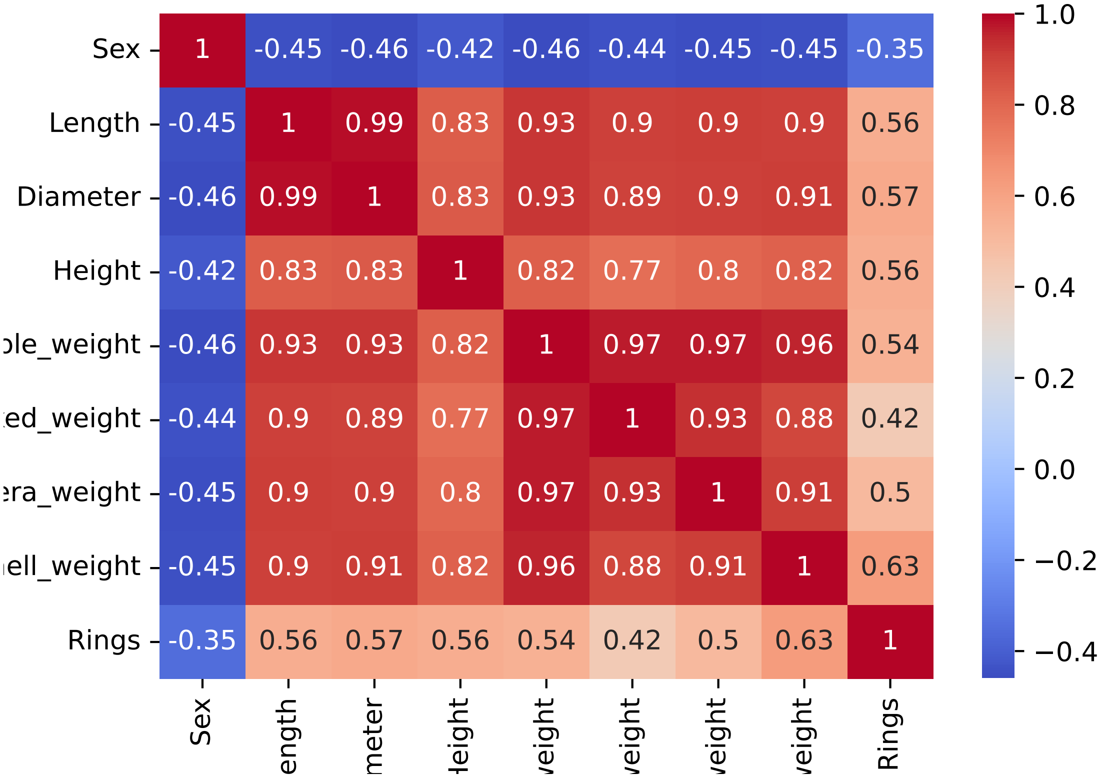
2 Exploratory Data Analysis

2.1 Develop a correlation map using a heatmap and discuss major observations

Correlation Matrix with all variables:



Correlation Matrix without the target variable:

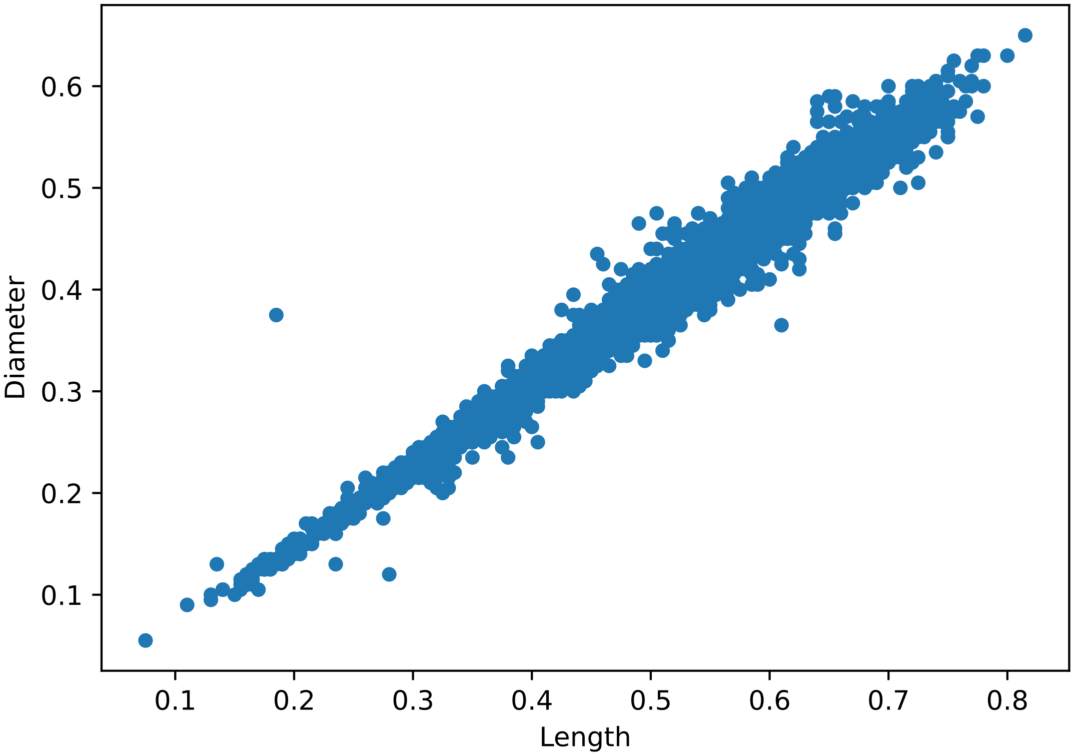


Whole Weight is almost linearly varying with all other features except ring age. Height has least linearity with remaining features. Ring age is most linearly proportional with shell weight, diameter and length. Ring age is least correlated with shucked weight.

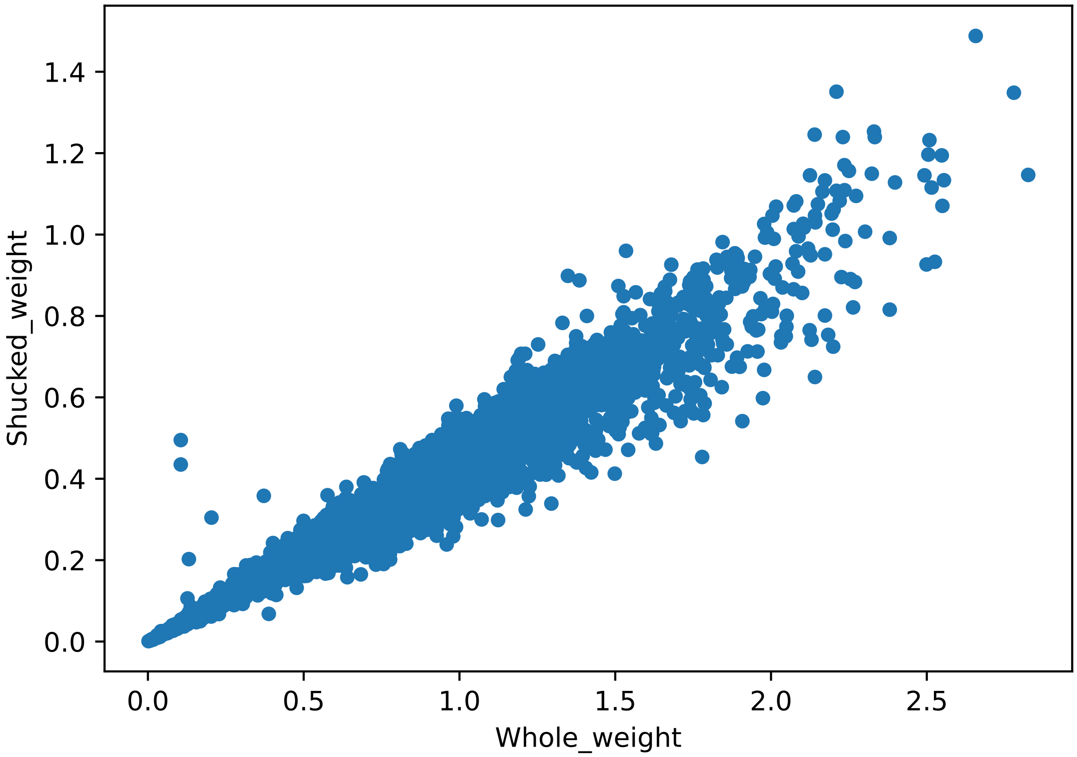
We can observe from the correlation matrix that Length - Diameter and Whole weight - Shucked weight has positive correlation as well as Sex - Height (-0.42) and Sex - Shucked Weight (-0.44) have least correlation.

2.1.1 Observations from the two most positively and negatively correlated features

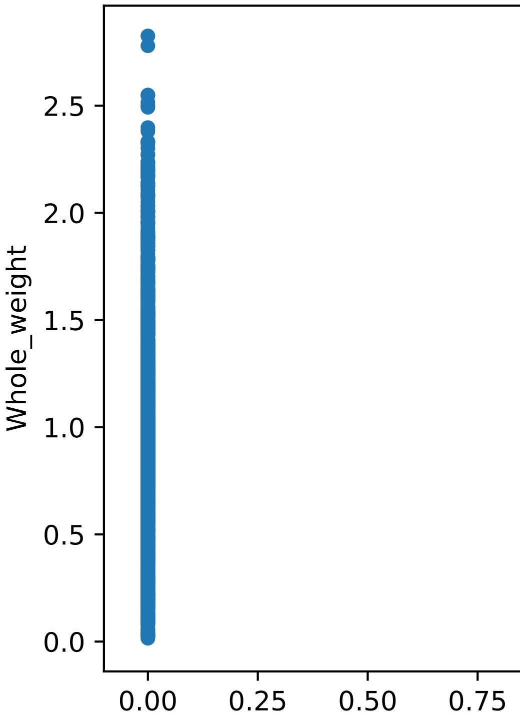
2.1.1.1 Observing Length Vs. Diameter and Whole_weight Vs. Shucked_weight



In the length vs diameter plot, we can observe from this plot the correlation is very strong and the trend is upward and data points are not too dispersed.

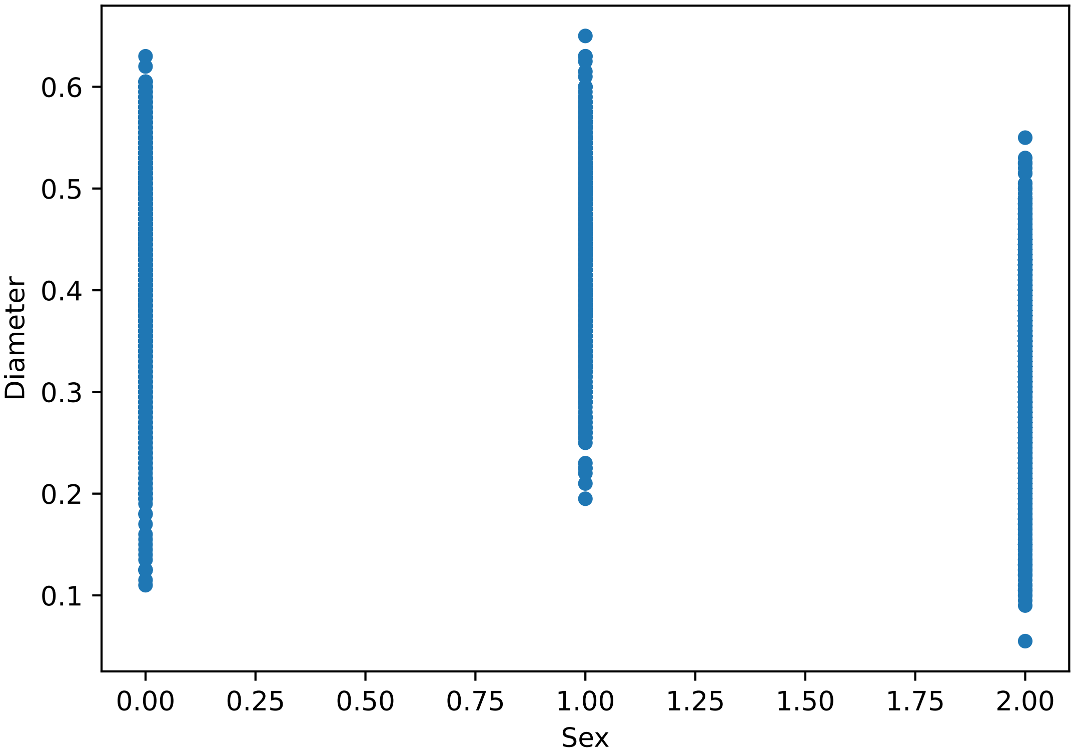


Same as the earlier plot we observed that whole weight vs shucked weight are strongly correlated with a upward trend but in this plot we can see that data points are bit dispersed when the whole weight is over 2gm.



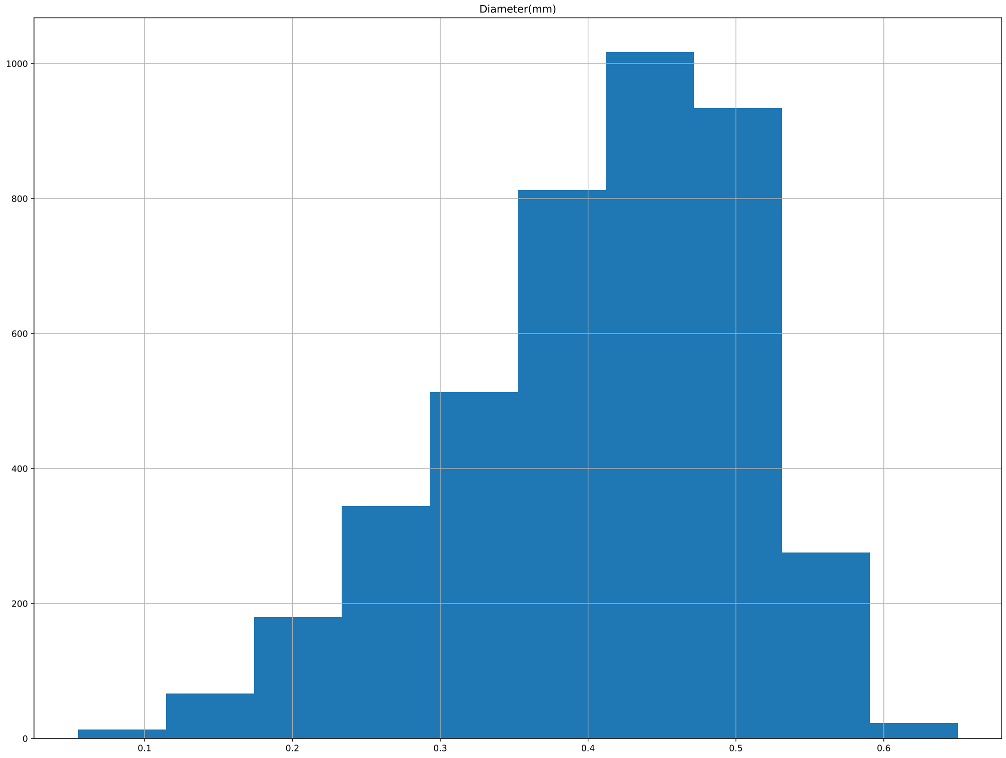
2.1.1.2 Observing Sex Vs. Whole weight and Sex Vs. Diameter

Among the least correlated features according the correlation matrix we choose sex vs whole weight plot to understand the relationship. Here we can see that as sex is categorized in three groups - male, female and infant as represented by 0 , 1 and 2 in the plot. We observe male abalones whole weight is wide spread above from 0 to 2.4gm (approx). Also we can almost say the same for female abalones but the infant ones mostly varies between 0 to 1.5mm.

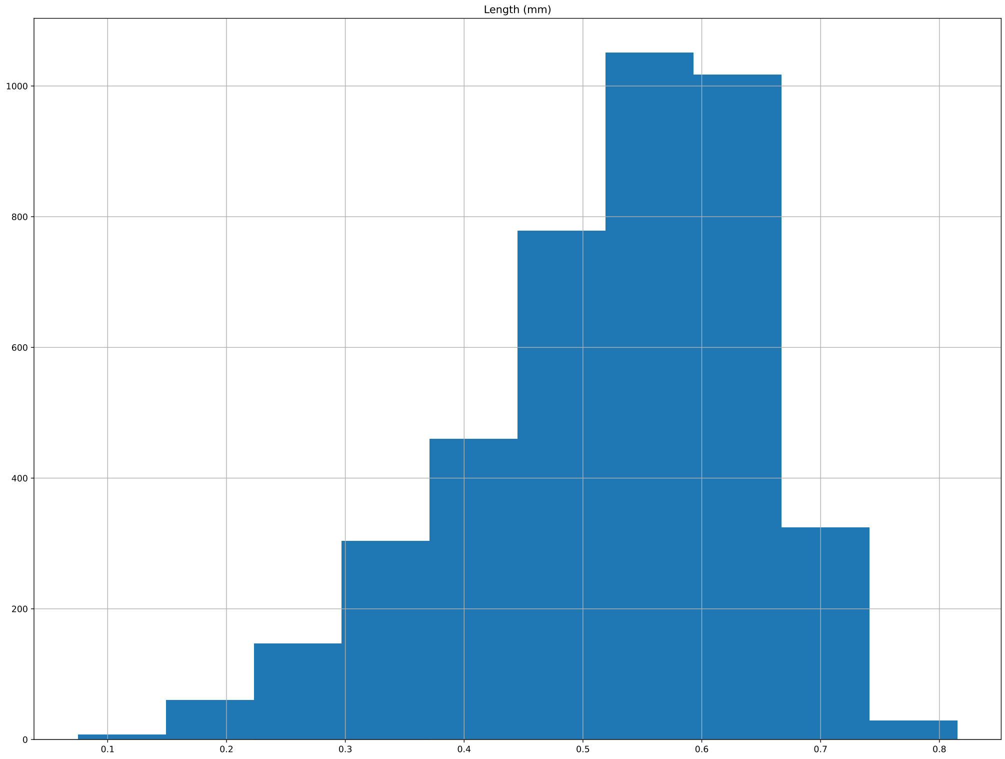


Another least correlated features according to the matrix are sex and diameter. In the plot we observe that male(0) abalones diameter are ranges mostly between 0.1mm and 0.6mm. Female(1) abalones diameter ranges from around 1.8mm 0.65mm. And the infant abalones mostly ranges from 0.1 mm to .55mm.

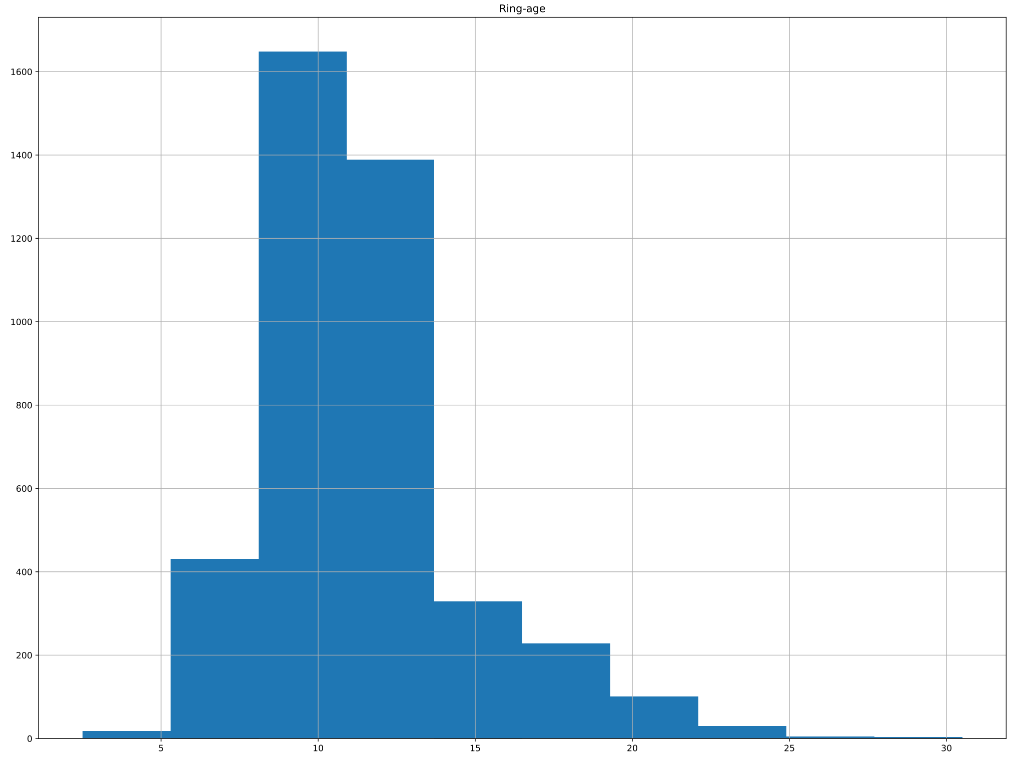
2.1.2 Observations from the histograms of the two most correlated features and the Ring-age



We observed from the plot that the data is left skewed or negatively skewed for the diameter feature. Most of the data are on the right side.



We observed from the plot that the data is left skewed or negatively skewed for the length feature. Most of the data are on the right side. We observe normal distribution for the both plots.



We observed from the plot that the data is right skewed or positively skewed for the ring age feature. Most of the data are on the left.

3 Experiment setup for Linear Regression Modelling

We have created 60/40 train/test split in this experiment. We have randomized the seed value based on the number of experiment. We have run 30 experiments in this study. For normalizing data set we have used scikit normalizer.

4 Linear Regression Modelling

After running the model 30 times with all the features to predict ring-age, we get the below result:

mean_rmse: 2.2347855732169999 std_rmse: 0.05060209021426619 mean_rsq: 0.5204818164392812 std_rsq: 0.02210463081444215

The mean RMSE(Root Mean Squared Error) value to the standard deviation (MSE to variance) and the standard deviation of RMSE (std_rmse) is a measure of how large the residuals are spread out. RMSE can range from 0 to positive infinity, so as it gets higher, it becomes harder to interpret how well the model is performing.

As higher the R-squared, the better the model fits the data. We observe a mean R-squared value of 0.52 which means a moderate uphill (positive) relationship between ring-age feature with all the other features.

In other word we can say 52% abalone ring age can be predicted using all the features. A low standard deviation of 0.022 indicates that R-squared values are mostly close to the mean.

4.1 Linear Regression Modelling with Normalization

After running the model 30 times with all the features with normalization to predict ring-age, we get the below result:

mean_rmse: 2.184666217489276 std_rmse: 0.04290735336508177 mean_rsq: 0.5419857466619026 std_rsq: 0.013262721603074073

When running the experiment with normalization we observe very less improvement in RMSE but with the new R-squred value we can say 54% of our data can be used to predict abalone age and the variation is also improved for R-squared.

After running the model 30 times with only two features - length and diameter to predict ring-age without normalization and with normaliztion, we get the below result:

Without Normalization: mean_rmse: 2.6374088456292957 std_rmse: 0.056172072178454525 mean_rsq: 0.3325955411414172 std_rsq: 0.016881975988073786

With Normalization: mean_rmse: 3.0290295819954864 std_rmse: 0.055220385377776585 mean_rsq: 0.11955200798148967 std_rsq: 0.0232855633057431

The above result tells us that after normalizing both RMSE and R-squared decreased, that indicates that normalization did not improve the model performance and data points are wide spread. Also prior normalizing the model only can predict 33% of abalone age using the length and diameter feature.

5 Conclusion

In this project we examine the Abalone dataset to determine the abalone age using linear regression machine learning technique using different sets of input features. In conclusion we notice that with more input features and appropriate normalization it would have been possible to determine more accurate age. However, our preliminary investigation indicates that linear regression is only suitable for a small set of input features and therefore the error estimates can be higher in comparing to use other advance machine learning techniques that we aim to use in the future.